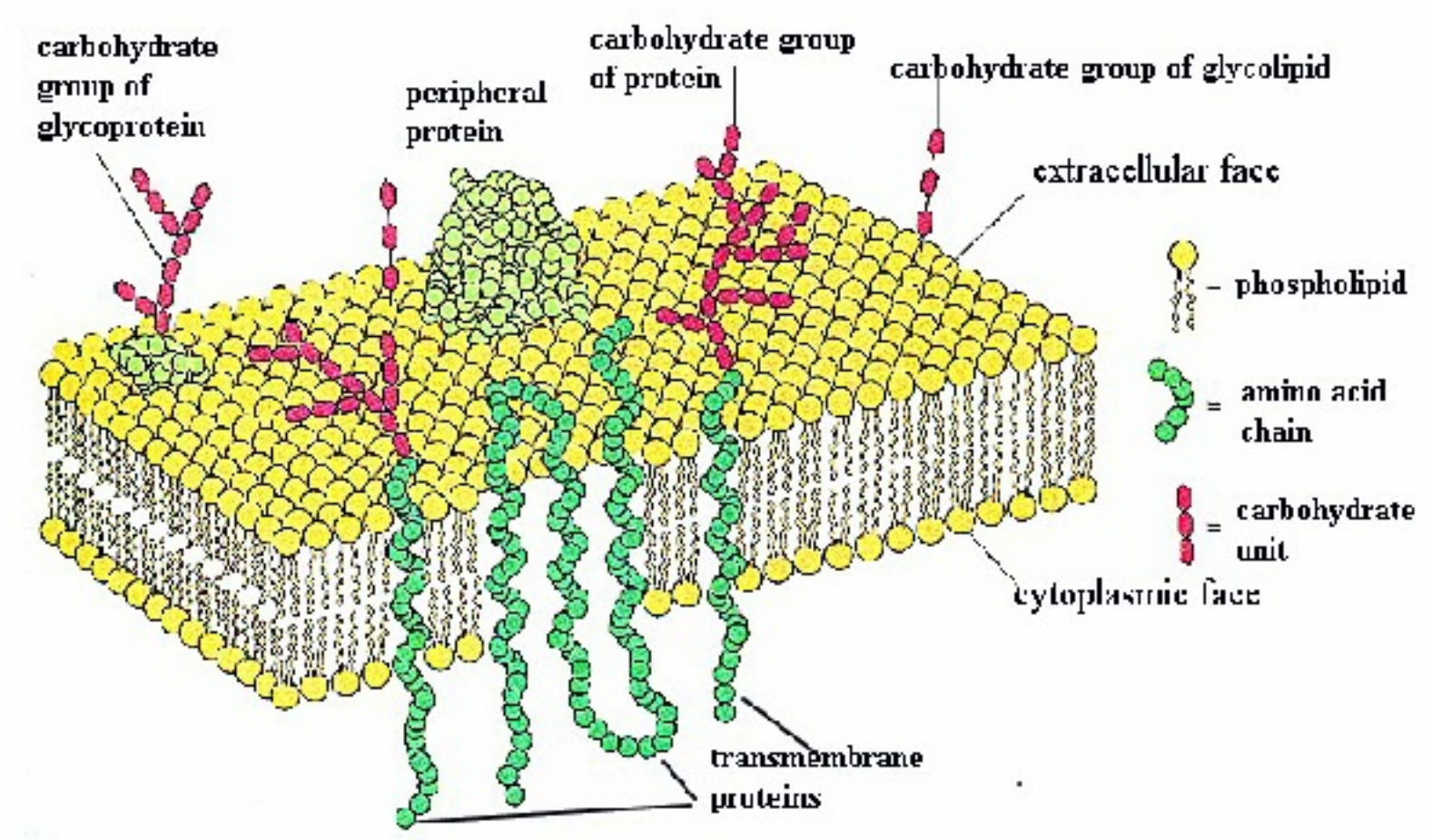
Lecture 3 Predicting Transmembrne proteins and coiled coils

Computational Aspects of Molecular Structure

Teresa Przytycka, PhD

Membrane and membrane proteins

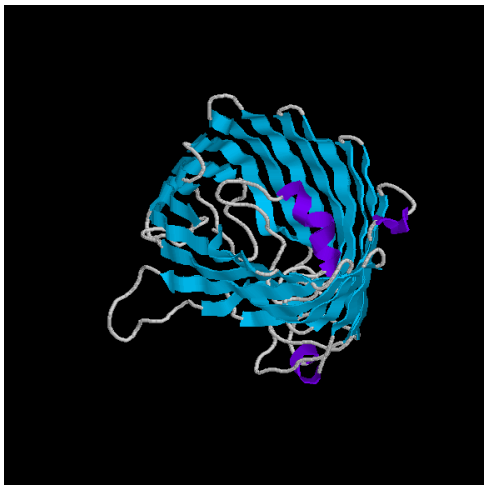


Importance of predicting of membrane proteins

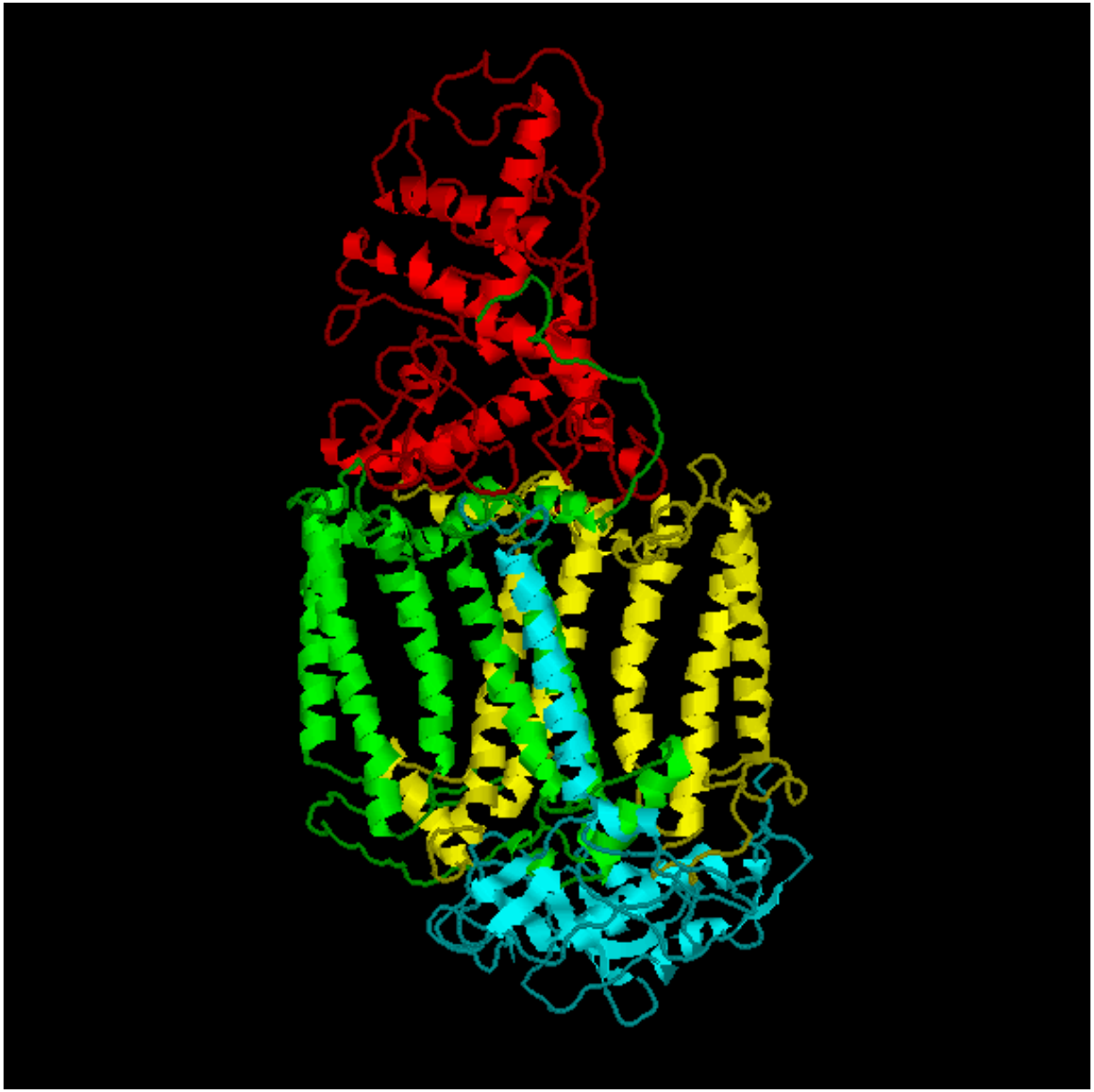
- About 30% of genomes encode for membrane proteins.
- Membrane proteins perform many important function: pores, ion channels, receptors.
- Only handful of membrane proteins is solved.
- Recognition algorithms for globular proteins do not work for membrane proteins.

Two classes of membrane proteins

- Helical bundle (ex. photosynthesis reaction center)
- beta-barrels (porins)



Photosynthesis reaction center



Membrane proteins are restrained by lipid environment

1. Membrane helices don’t make hydrogen bonds with solvent.
2. Membrane beta-barrels pass water molecules “ thorough the pore thus are hydrophobic outside and hydrophilic inside
3. Predicting fold of transmembrane proteins is potentially easier than water solvable proteins due to severely restricted way in which a protein can be embedded in the membrane

Why methods for water soluble proteins do not work for trans membrane proteins

- not enough data to collect good statistics.
- most transmembrane proteins are helical bundles – so the recognition problem is vary specific.
- The transmembrane beta-barrels have even number of strands.
- transmembrane protein will tolerate substantial drift in sequence without change in structure (no much help with profile methods).

Early methods

- Kyte and Doolittle : hydropathy plots to predict transmembrane helices: *Transmembrane helices are buried in the non-polar phase of the lipid membrane whilst other part (loops) exist in more polar solution.*
- Heijne: positive inside rule *positively charged residues (Arg, Lys) tend to be much more frequent in non-translocated regions as compared to translocated regions.*

Hydrophobicity plots

- Most methods for predicting transmembrane helices start by computing hydropathy plot.
- There are many hydrophobicity scales. Some computed from experimental solution study of free energy transfer from aqueous solution to that that mimics membrane, some use crystallographic data.

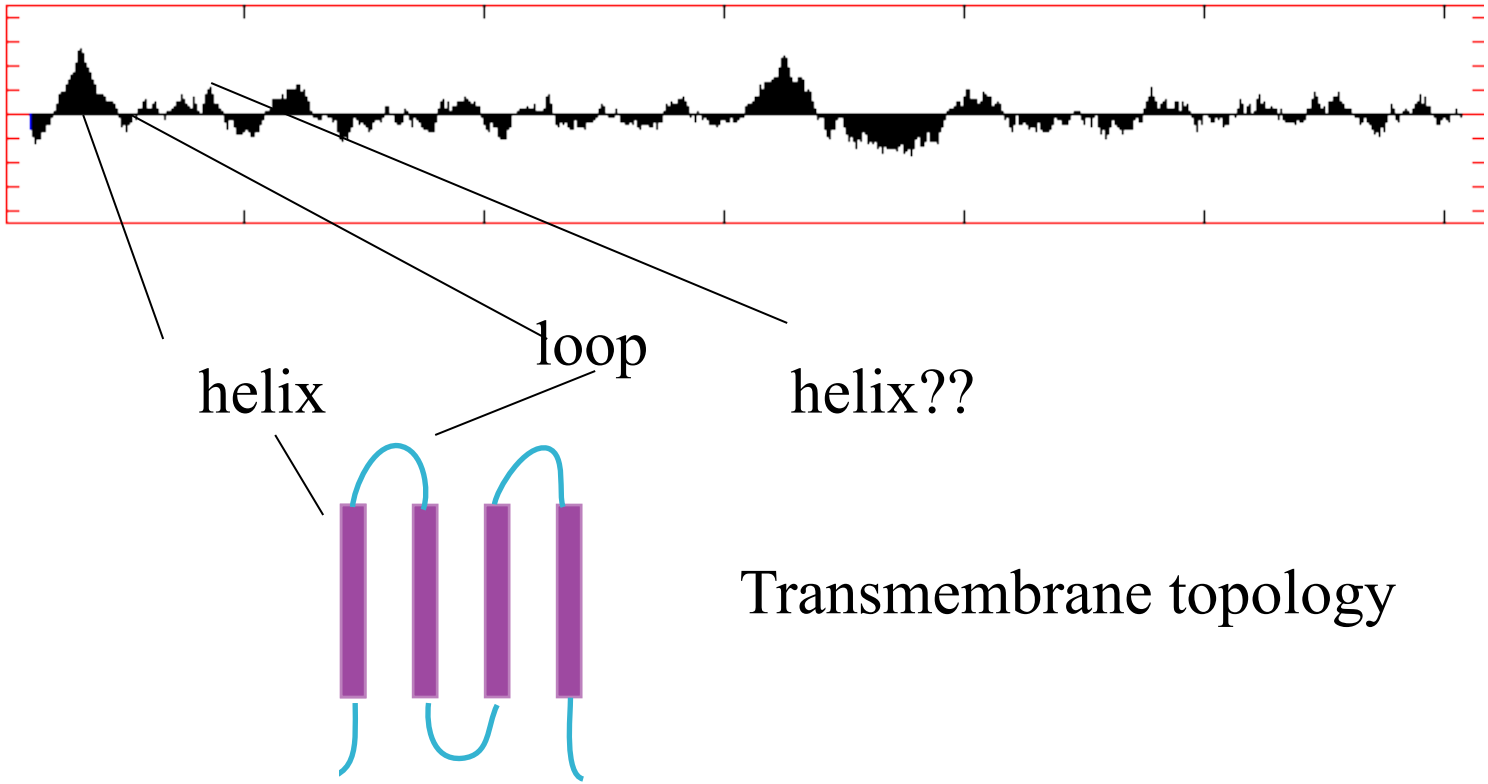
Kyte Doolittle scale (1982)

G	-0.4	**Q**	-3.5	**S**	-0.8	**Y**	-1.3
A	1.8	**K**	-3.9	**T**	-0.7	**W**	-0.9
V	4.2	**H**	-3.2	**D**	-3.5	**C**	2.5
L	3.8	**R**	-4.5	**E**	-3.5	**M**	1.9
I	4.5	**F**	2.8	**N**	-3.5	**P**	-1.6

Hydropathy plot

Slide a window and for each residuum include the contribution of neighboring residues as followsL

$$H(a_i) = \Sigma_{i-k<l<i+k} h(a_l)$$



Transmembrane topology

Sample prediction protocol

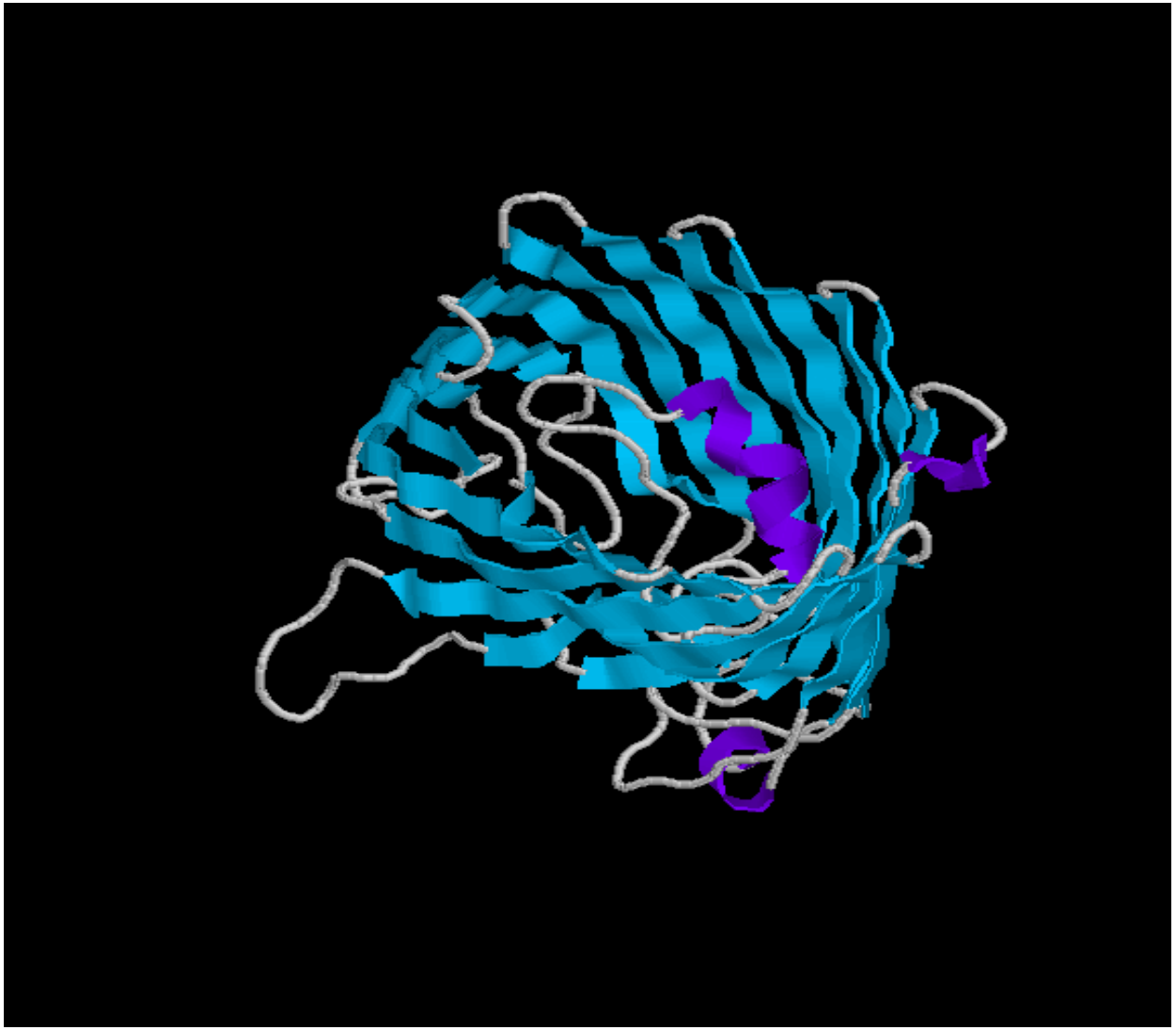
(TOP-PRED Sipos, Heijne)

- Construct hydropathy plot
- Identify "certain" helices (peeks above "upper" cut-off)
- Identify "putative" helices (peeks above "lower"cut-off but below upper cut-off)
- Construct all possible topologies that include all "sure" helices and include or exclude putative segments (we are not concern with the helix position in the membrane but only in finding the helices)
- For each possible topology compute $\Delta+$ = the difference between the number of Arg+Lys between the two sides of the structure (exclude long loops)
- Chose the structure with largest $\Delta+$

Difficulties and newer methods

- The "positive inside" rule is often disturbed by globular domains in the loops.
- Signal peptides are also stretches of hydrophobic residues so we need to recognize what is a transmarine helix and what a signal peptide
- New methods makes use of Hidden Markov Models and will be discussed later.

Beta-barrel (porin)



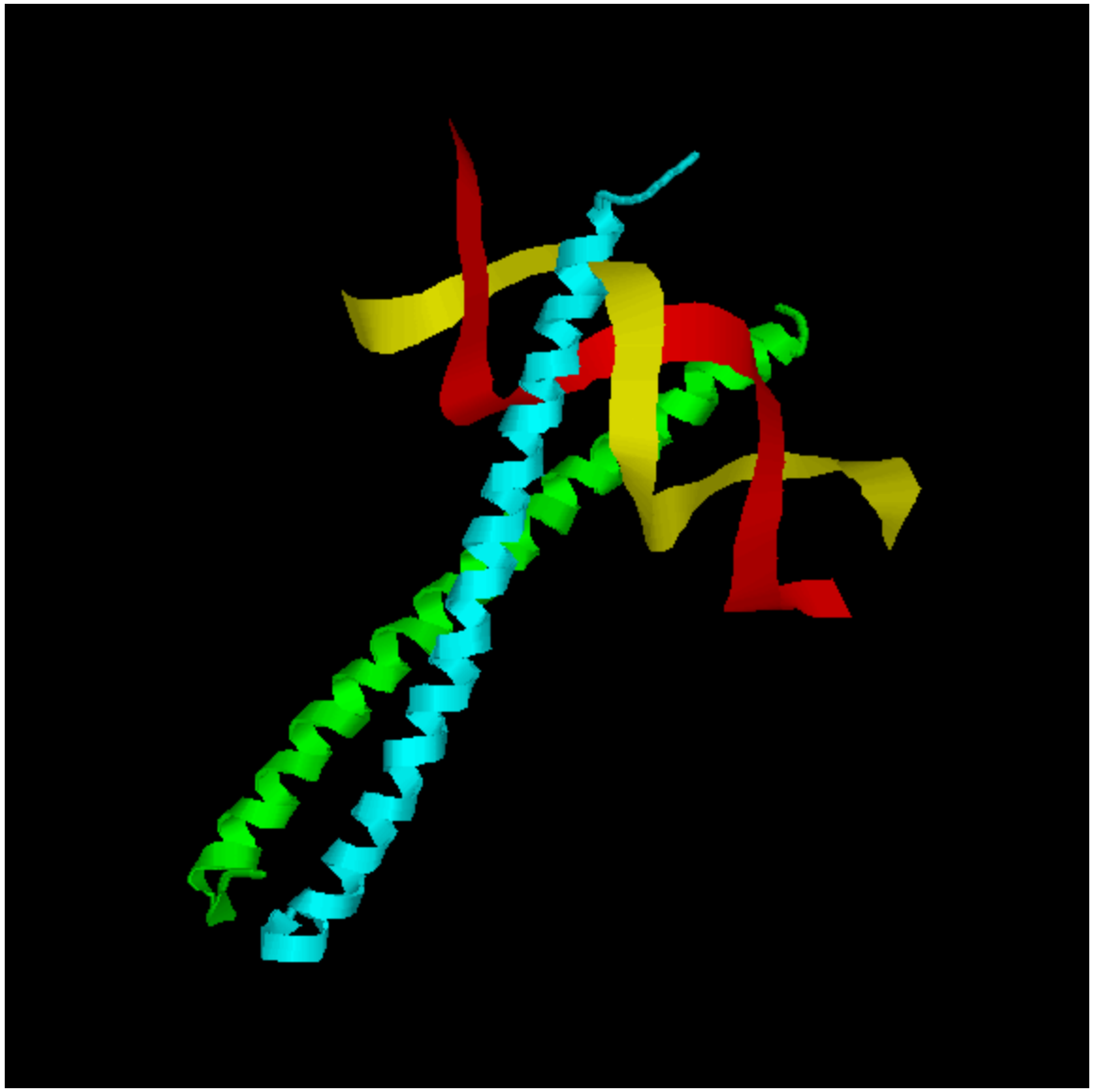
Properties of β-barrel membrane proteins

- The helix prediction methods cannot be used for strands
- In the transmembrane strands every second residue is hydrophobic and faces the lipid.
- Sided hydrophobicity profile:

  $$H(i) = \tfrac{1}{4}(h(i-2)+h(i) + h(i+2)+h(i+4))$$

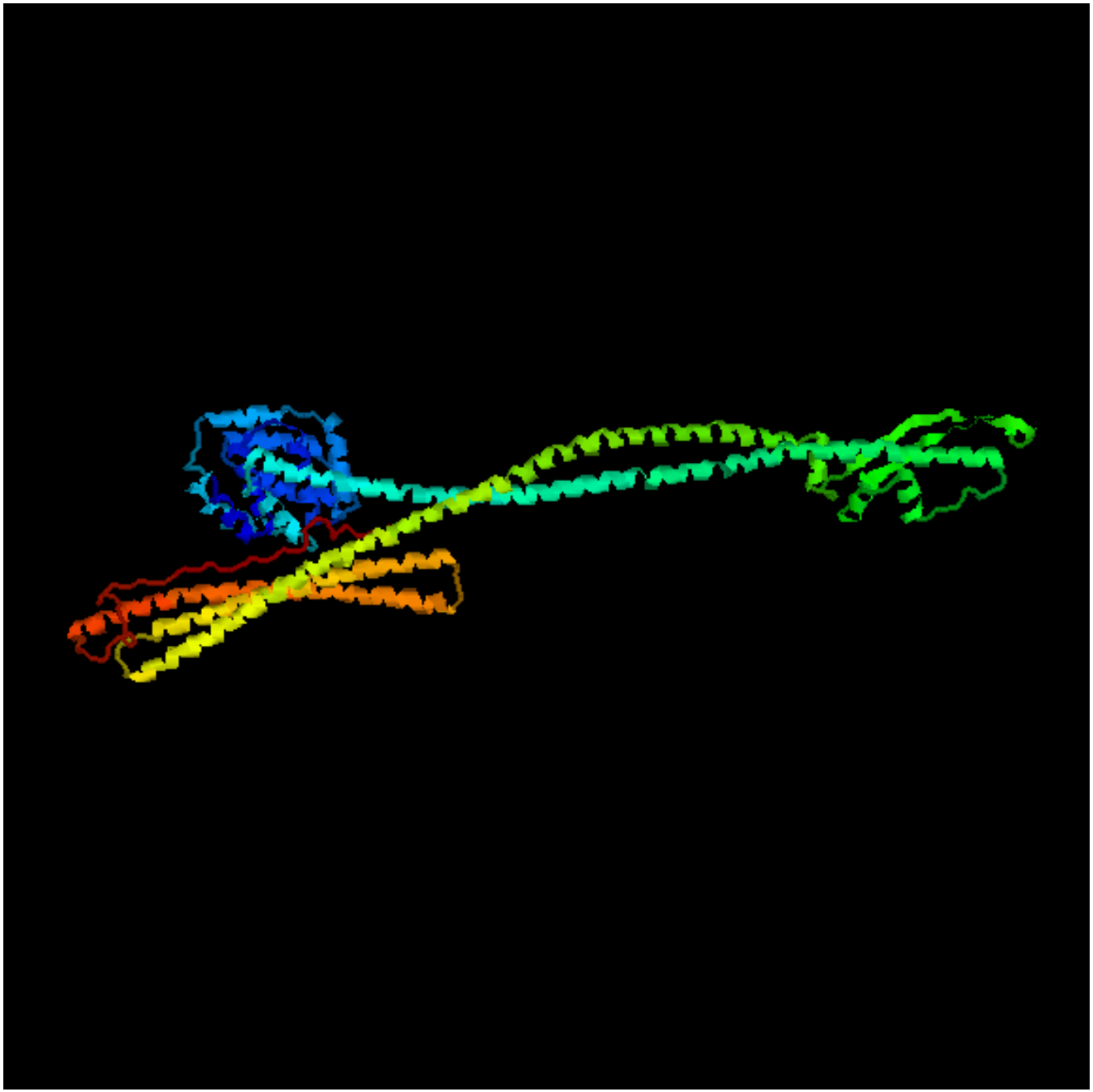
  It helps to find the "every second hydrophobic" pattern)
- Strands are frequently flanked by aromatic residues (Phe, Try, Trp)

Coiled Coil



- Two or three helices twisted together
- Usually quite long (hard to crystallize)
- Many structural proteins

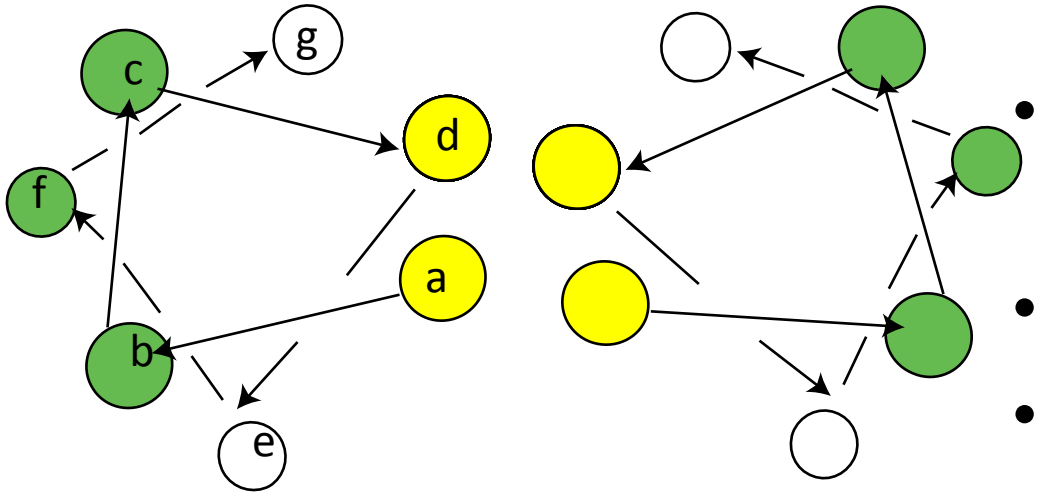
lucine zipper



Coil-coiled proteins:

- Lucine zipper (DNA binding protein)
- Involved in tRNA synthesis
- Membrane fusion proteins (play a role in how HIV and other viruses enter a cell)
- Muscle proteins

Heptads repeat - characteristic repeating pattern



- a,b,c,d,e,f,g – repeating 7 residues (heptads)
- a,d – tightly packed in hydrophobic core (large)
- b,c,f – frequently charged
- Profile analysis shows that preference for residues to be in particular position of the repeats.

Coiled coil has been synthesized de novo

protein1 AQLEKELQAQLEKELQAQL

protein2 AQLKKKLQAQLKKKLQAQL

Separately both proteins are random coils, together form stable coiled coil

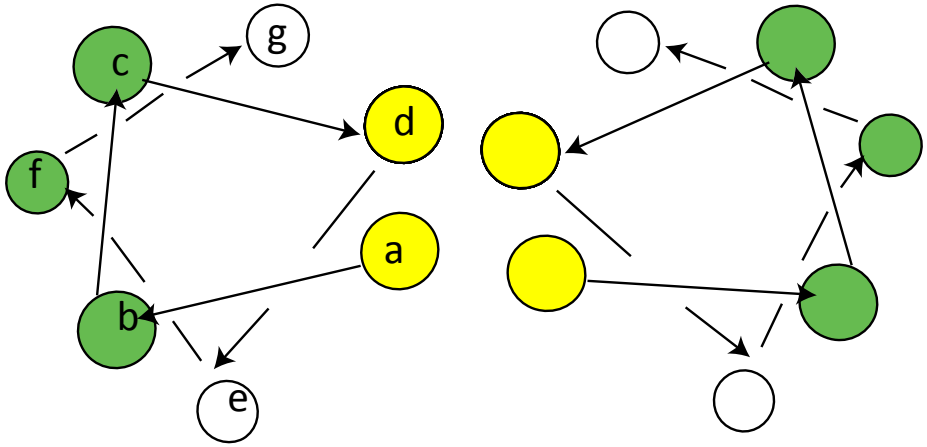
E - Glutamic acid – positively charged

K – Lysin negatively charged

L – Leucine – large hydrophobic

Simple coiled coil recognition algorithm

	a	b	c	d	e	f	g
L							
A							
…							



Compute profile :

$$\frac{\%\ \text{lucine in cc at pos a}}{\%\ \text{lucine in data base}}$$

For each position (a-g) compute the propensity of a given amino acid to be at the given position

In the given input sequence look for stretches of high propensity residua (at last 28 res. long)

(Parry –82, Lupas, van Dyke, Stock-91 (NewCoil), Fischetti, Landau, Schmidt, Sellers-93)

Problem – high false positive rate (2/3)

Towards more advanced methods

- The previous coiled-coil recognition method was similar to C-F secondary structure prediction algorithm
- Next step: a GOR-like approach: compute probability that given residuum is part of coiled-coil in the context if a window of its neighbors.

Probabilistic framework

- Given a subsequence $z=r_1,\ldots, r_{28}$ what is the probability` that it is a coiled coil (CC) (28=minimum coiled coil length)
- Let $X= R_1,\ldots, R_{28}$ be a random sequence from a data base then:

$$P[z \text{ is CC}] = P[X \text{ is CC}| X=z] =$$

$$P[X \text{ is CC } \& \ X =Z] / P[X=Z]=$$

$$P[(X \text{ is CC}) \ \& \ (R_1= r_1)\&\ldots\& \ (R_{28}= r_{28})] /$$

$$P[(R_1= r_1)\&\ldots\& \ (R_{28}= r_{28})]$$

- The data base does not contain enough data to estimate the above probabilities based on frequencies of occurrences (same argument as in GOR algorithm).
- If we assume that positions are independent we get the propensity table approach.
- Idea: assume dependence of some residues pairs and keep the rest independent (compare to GOR).

Pair Coil algorithm

Berger (1995)

- Main idea: explore correlations between i and i+1, i+3, i+4 but – still not enough data to approximate frequencies.
- Instead pair-wise dependencies ware used (like GOR)
- Testing Pair Cols : no false positive
- Some coiled coils have been missed. Subsequently LearnCoil-Histidine Kinase and LearnCoil-VMF ware written (Singh 1998,1999) to predict special families of coiled coils